加微信：642945106 发送“ 赠送 ”领取赠送精品课程
发数字“ 2 ”获取众筹列表

01 | 前因后果：为什么说数据中台是大数据的下一站?

2020-03-30 郭忆

数据中台实战课 进入课程 >

讲述：郭忆
时长 21:07 大小 19.35M

你好，我是郭忆。

“数据中台”无疑是今年大数据圈最火的词，如果你关注数据相关的行业会议，但凡有数据中台相关的主题，人员都会爆满。去年 5 月，我作为演讲嘉宾参加了由 ITPUB 主办的中国数据库大会，一个 100 人的“数据中台”场次，最后涌进来 200 多人，前排地下、走廊、过道到处都挤满了人，还有很多人因为挤不进来在外面看直播，数据中台的火爆程度可见一斑。

除了支撑集团的大数据建设，我的团队还提供 To B 的企业服务，因此我也有机会接触到一些正在做数字化转型的传统企业。从 2018 年末开始，原先市场上各种关于大数据平台的招

标突然不见了，取而代之的是数据中台项目，建设数据中台俨然成为传统企业数字化转型的首选，甚至不少大数据领域的专家都认为，数据中台是大数据的下一站。

那么为什么数据中台被认为是大数据的下一站呢？它与你之前遇到的数据仓库、数据湖、大数据平台又有什么区别？

今天这节课，我想带着这个问题，**与你深入大数据的发展历史，先从数据仓库的出现讲起，途径数据湖，再到大数据平台，**因为这样，你才能理解大数据发展的每个阶段遇到的问题，从而深入理解数据中台在大数据发展中的历史定位。

启蒙时代：数据仓库的出现

商业智能（Business Intelligence）诞生在上个世纪 90 年代，它是将企业已有的数据转化为知识，帮助企业做出经营分析决策。比如在零售行业的门店管理中，如何使得单个门店的利润最大化，我们就需要分析每个商品的销售数据和库存信息，为每个商品制定合理的销售采购计划，有的商品存在滞销，应该降价促销，有的商品比较畅销，需要根据对未来销售数据的预测，进行提前采购，这些都离不开大量的数据分析。

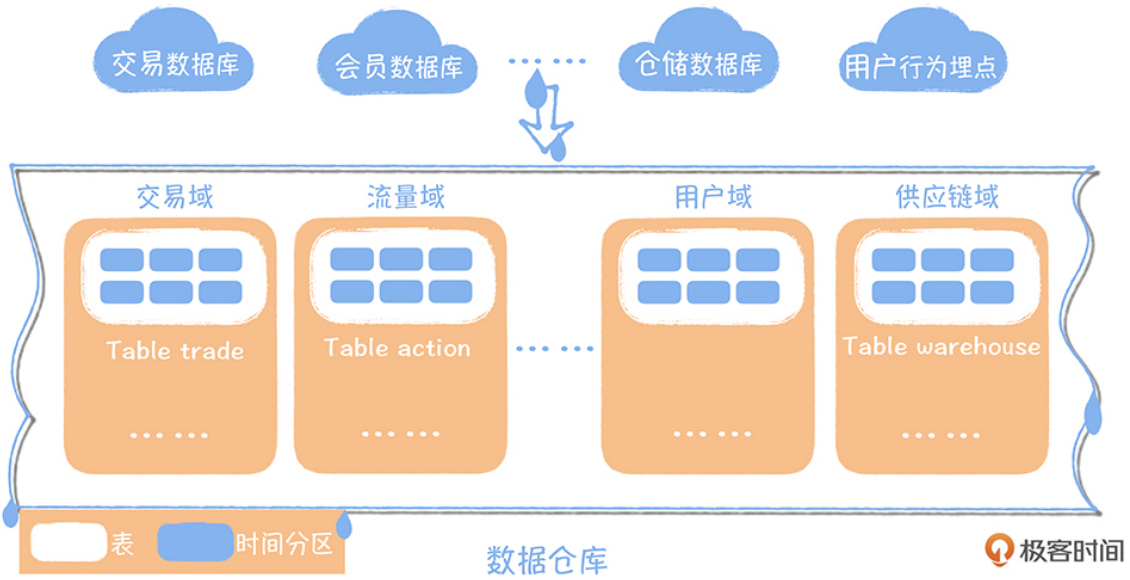
而数据分析需要聚合多个业务系统的数据，比如需要集成交易系统的数据，需要集成仓储系统的数据等等，同时需要保存历史数据，进行大数据量的范围查询。传统数据库面向单一业务系统，主要实现的是面向事务的增删改查，已经不能满足数据分析的场景，**这促使数据仓库概念的出现。**

在 1991 年出版的《Building the Data Warehouse》中，数据仓库之父比尔·恩门（Bill Inmon）首次给出了数据仓库的完整定义，他认为：

> 数据仓库是在企业管理和决策中面向主题的、集成的、与时间相关的，不可修改的数据集合。

为了帮你理解数据仓库的四要素，我举个电商的例子。

在电商场景中，有一个数据库专门存放订单的数据，另外一个数据库存放会员相关的数据。构建数据仓库，首先要把不同业务系统的数据同步到一个统一的数据仓库中，然后按照主题域方式组织数据。



主题域是业务过程的一个高层次的抽象，像商品、交易、用户、流量都能作为一个主题域，**你可以把它理解为数据仓库的一个目录。**数据仓库中的数据一般是按照时间进行分区存放，一般会保留 5 年以上，每个时间分区内的数据都是追加写的方式，对于某条记录是不可更新的。

除了这个概念之外，我还要提一下他和金博尔（Kimball）共同开创的数仓建模的设计方法，这个方法对于后来基于数据湖的现代数据仓库的设计有重要的意义，所以你有必要了解。

防止断更 请务必加
首发微信：17161543665

恩门提出的建模方法自顶向下（这里的顶是指数据的来源，在传统数据仓库中，就是各个业务数据库），基于业务中各个实体以及实体之间的关系，构建数据仓库。

比如，在一个最简单的买家购买商品的场景中，按照恩门建模的思维模式，首先你要理清这个业务过程中涉及哪些实体。买家、商品是一个实体，买家购买商品是一个关系。所以，模型设计应该有买家表，商品表，和买家商品交易表三个模型。

用户ID	用户账号	用户账号余额	创建时间	最近修改时间
1	XiaoMing	1000	2019-1-10 10：20：20	2019-01-11 10：59：12
2	XiaoLi	500	2019-1-10 10：30：21	2019-01-11 11：20：11

极客时间

买家表

商品ID	商品名称	商品规格	商品库存	创建时间	最近修改时间
1	沐浴露	500ML	2000	2018-11-10 10：20：11	2019-01-11 10：59：12
2	婴儿车	0～3岁	100	2018-12-10 11：20：11	2019-01-11 08：22：11

商品表

用户ID	商品ID	交易金额	交易时间
1	1	20	2019-01-11 10：59：12
2	2	500	2019-01-11 08：22：11

买家商品交易表

金博尔建模与恩门正好相反，是一种自底向上的模型设计方法，从数据分析的需求出发，拆分维度和事实。那么用户、商品就是维度，库存、用户账户余额是事实。

用户ID	用户账号	创建时间	最近修改时间
1	XiaoMing	2019-1-10 10：20：20	2019-01-11 10：59：12
2	XiaoLi	2019-1-10 10：30：21	2019-01-11 11：20：11

用户维度表

商品ID	商品名称	商品规格	创建时间	最近修改时间
1	沐浴露	500ML	2019-1-10 10：20：20	2019-01-11 10：59：12
2	婴儿车	0～3岁	2019-1-10 10：30：21	2019-01-11 11：20：11

商品维度表

用户ID	账户余额	交易金额	最近修改时间
1	1000	20	2019-01-11 10：59：12
2	500	50	2019-01-11 11：20：11

账户余额事实表

商品ID	商品库存	交易数量	交易时间
1	2000	1	2019-01-11 10：59：12
2	100	1	2019-01-11 11：20：11

商品库存事实表

这两种方法各有优劣，恩门建模因为是从数据源开始构建，构建成本比较高，适用于应用场景比较固定的业务，比如金融领域，冗余数据少是它的优势。金博尔建模由于是从分析场景

出发，适用于变化速度比较快的业务，比如互联网业务。**由于现在的业务变化都比较快，所以我更推荐金博尔的建模设计方法。**

传统数据仓库，第一次明确了数据分析的应用场景应该用单独的解决方案去实现，不再依赖于业务的数据库。在模型设计上，提出了数据仓库模型设计的方法论，为后来数据分析的大规模应用奠定了基础。但是进入互联网时代后，传统数据仓库逐渐没落，一场由互联网巨头发起的技术革命催生了大数据时代的到来。

技术革命：从 Hadoop 到数据湖

进入互联网时代，有两个最重要的变化。

> 一个是数据规模前所未有，一个成功的互联网产品日活可以过亿，就像你熟知的头条、抖音、快手、网易云音乐，每天产生几千亿的用户行为。传统数据仓库难于扩展，根本无法承载如此规模的海量数据。
>
> 另一个是数据类型变得异构化，互联网时代的数据除了来自业务数据库的结构化数据，还有来自 App、Web 的前端埋点数据，或者业务服务器的后端埋点日志，这些数据一般都是半结构化，甚至无结构的。传统数据仓库对数据模型有严格的要求，在数据导入到数据仓库前，数据模型就必须事先定义好，数据必须按照模型设计存储。

所以，数据规模和数据类型的限制，导致传统数据仓库无法支撑互联网时代的商业智能。

而以谷歌和亚马逊为代表的互联网巨头率先开始了相关探索。从 2003 年开始，互联网巨头谷歌先后发表了 3 篇论文：《The Google File System》《MapReduce：Simplified Data Processing on Large Clusters》《Bigtable：A Distributed Storage System for Structed Data》，这三篇论文奠定了现代大数据的技术基础。它们提出了一种新的，面向数据分析的海量异构数据的统一计算、存储的方法。关于这三篇论文，在这里我们不做深入的解读，如果对实现技术感兴趣的话，也可以查看我在文末提供的链接。

但 2005 年 Hadoop 出现的时候，大数据技术才开始普及。你可以把 Hadoop 认为是前面三篇论文的一个开源实现，我认为 Hadoop 相比传统数据仓库主要有两个优势：

1. 完全分布式，易于扩展，可以使用价格低廉的机器堆出一个计算、存储能力很强的集群，满足海量数据的处理要求；

2. 弱化数据格式，数据被集成到 Hadoop 之后，可以不保留任何数据格式，数据模型与数据存储分离，数据在被使用的时候，可以按照不同的模型读取，满足异构数据灵活分析的需求。

随着 Hadoop 技术日趋成熟，2010 年，Pentaho 创始人兼 CTO James Dixon 在纽约 Hadoop World 大会上提出了数据湖的概念，他提到：

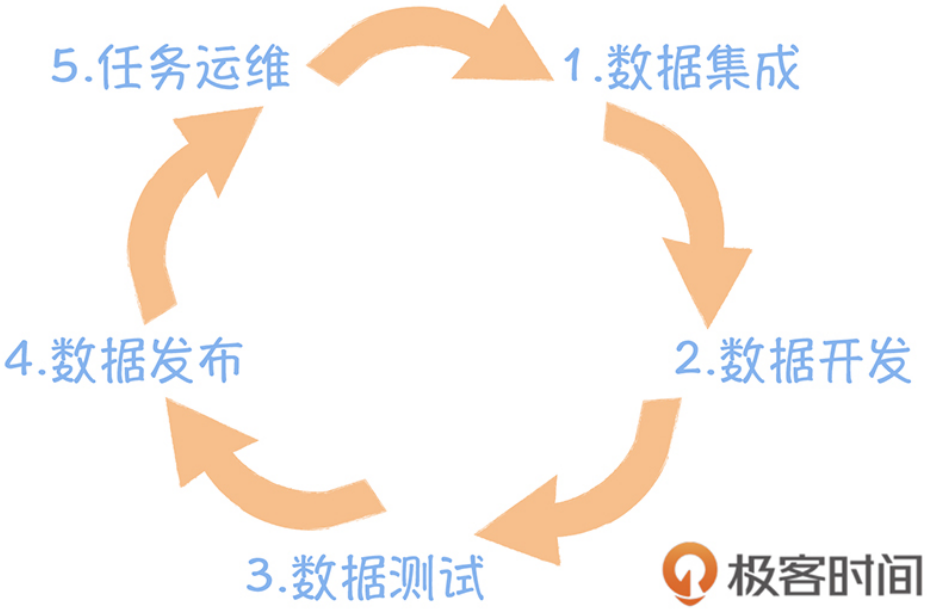
> 数据湖（Data Lake）是一个以原始格式存储数据的存储库或系统。

数据湖概念的提出，我认为是 Hadoop 从开源技术走向商业化成熟的标志。企业可以基于 Hadoop 构建数据湖，将数据作为一种企业核心资产。

数据湖拉开了 Hadoop 商用化的大幕，但是一个商用的 Hadoop 包含 20 多种计算引擎，数据研发涉及流程非常多，技术门槛限制了 Hadoop 的商用化进程。那么如何让数据的加工像工厂一样，直接在设备流水线上完成呢？

数据工厂时代：大数据平台兴起

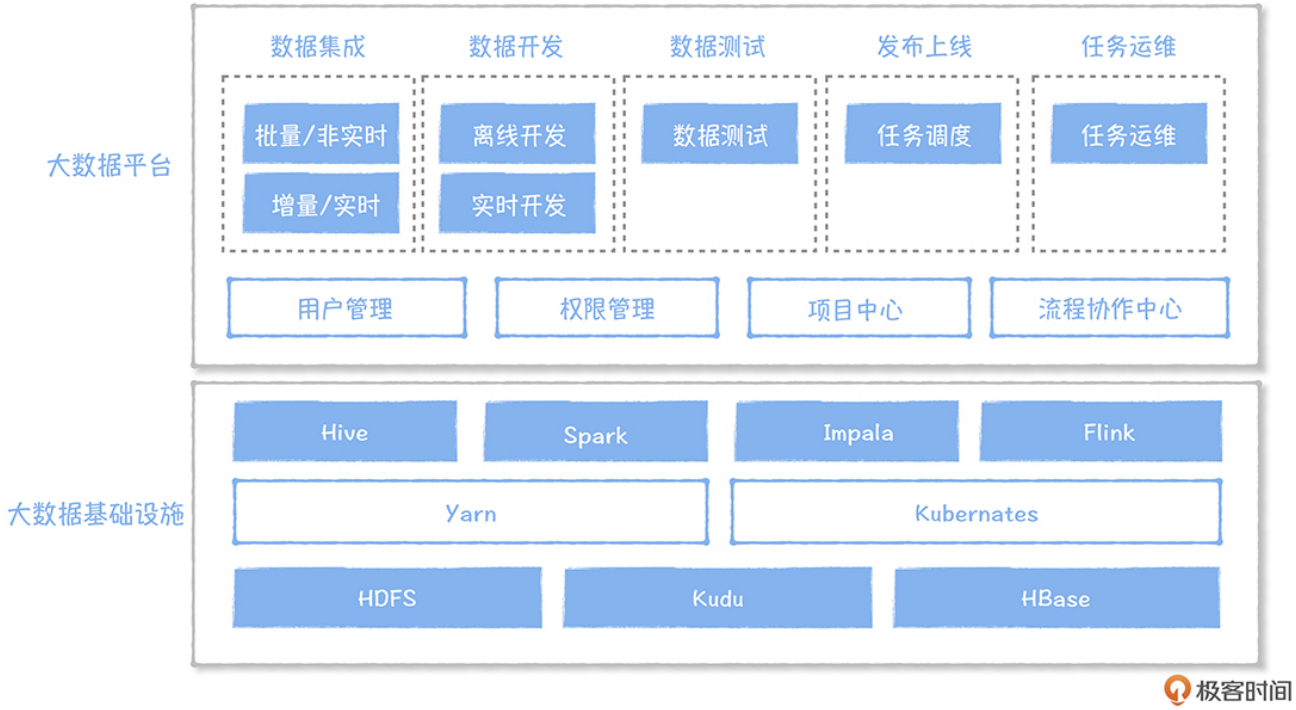
对于一个数据开发，在完成一项需求时，常见的一个流程是首先要把数据导入到大数据平台中，然后按照需求进行数据开发。开发完成以后要进行数据验证比对，确认是否符合预期。接下来是把数据发布上线，提交调度。最后是日常的任务运维，确保任务每日能够正常产出数据。



如此繁杂的一个工作流程，如果没有一个高效的平台作为支撑，就跟写代码没有一个好用的 IDE， 用文本编辑器写代码一样，别人完成十个需求，你可能连一个需求都完成不了，效率异常低下，根本无法大规模的应用。

提出大数据平台的概念，就是为了提高数据研发的效率，降低数据研发的门槛，让数据能够在一个设备流水线上快速地完成加工。

> 大数据平台是面向数据研发场景的，覆盖数据研发的完整链路的数据工作台



大数据平台按照使用场景，分为数据集成、数据开发、数据测试……任务运维，大数据平台的使用对象是数据开发。大数据平台的底层是以 Hadoop 为代表的基础设施，分为计算、资源调度和存储。

Hive、Spark、Flink、Impala 提供了大数据计算引擎：

> Hive、Spark 主要解决离线数据清洗、加工的场景，目前，Spark 用得越来越多，性能要比 Hive 高不少；
>
> Flink 主要是解决实时计算的场景；
>
> Impala 主要是解决交互式查询的场景。

这些计算引擎统一运行在一个称为 Yarn 的资源调度管理框架内，由 Yarn 来分配计算资源。目前最新的研究方向中也有基于 Kubernetes 实现资源调度的，例如在最新的 Spark 版本（2.4.4）中，Spark 已经能够运行在 Kubernetes 管理的集群上，这样的好处是可以实现在线和离线的资源混合部署，节省机器成本。

数据存储在 HDFS、Kudu 和 HBase 系统内。HDFS 不可更新，主要存全量数据，HBase 提供了一个可更新的 KV，主要存一些维度表，Kudu 提供了实时更新的能力，一般用在实时数仓的构建场景中。

大数据平台像一条设备流水线，经过大数据平台的加工，原始数据变成了指标，出现在各个报表或者数据产品中。随着数据需求的快速增长，报表、指标、数据模型越来越多，找不到数据，数据不好用，数据需求响应速度慢等问题日益尖锐，成为阻塞数据产生价值的绊脚石。

数据价值时代：数据中台崛起

时间到了 2016 年前后，互联网高速发展，背后对数据的需求越来越多，数据的应用场景也越来越多，有大量的数据产品进入到了我们运营的日常工作，成为运营工作中不可或缺的一部分。在电商业务中，有供应链系统，供应链系统会根据各个商品的毛利、库存、销售数据以及商品的舆情，产生商品的补货决策，然后推送给采购系统。

大规模数据的应用，也逐渐暴露出现一些问题。

业务发展前期，为了快速实现业务的需求，烟囱式的开发导致企业不同业务线，甚至相同业务线的不同应用之间，数据都是割裂的。两个数据应用的相同指标，展示的结果不一致，导致运营对数据的信任度下降。如果你是运营，当你想看一下商品的销售额，发现两个报表上，都叫销售额的指标出现了两个值，你的感受如何？你第一反应肯定是数据算错了，你不敢继续使用这个数据了。

数据割裂的另外一个问题，就是大量的重复计算、开发，导致的研发效率的浪费，计算、存储资源的浪费，大数据的应用成本越来越高。

> 如果你是运营，当你想要一个数据的时候，开发告诉你至少需要一周，你肯定想是不是太慢了，能不能再快一点儿？

如果你是数据开发，当面对大量的需求的时候，你肯定是在抱怨，需求太多，人太少，活干不完。

如果你是一个企业的老板，当你看到每个月的账单成指数级增长的时候，你肯定觉得这也太贵了，能不能再省一点，要不吃不消了。

这些问题的根源在于，数据无法共享。2016 年，阿里巴巴率先提出了“数据中台”的口号。**数据中台的核心，是避免数据的重复计算，通过数据服务化，提高数据的共享能力，赋能数据应用。**之前，数据是要啥没啥，中间数据难于共享，无法积累。现在建设数据中台之后，要啥有啥，数据应用的研发速度不再受限于数据开发的速度，一夜之间，我们就可以根据场景，孵化出很多数据应用，这些应用让数据产生价值。

课堂总结

现在，回到我们本节课的题目：为什么说数据中台是大数据的下一站？ 在我看来，有这样几个原因：

1. 数据中台构建于数据湖之上，具备数据湖异构数据统一计算、存储的能力，同时让数据湖中杂乱的数据通过规范化的方式管理起来。
2. 数据中台需要依赖大数据平台，大数据平台完成了数据研发的全流程覆盖，数据中台增加了数据治理和数据服务化的内容。
3. 数据中台借鉴了传统数据仓库面向主题域的数据组织模式，基于维度建模的理论，构建统一的数据公共层。

总的来说，数据中台吸收了传统数据仓库、数据湖、大数据平台的优势，同时又解决了数据共享的难题，通过数据应用，实现数据价值的落地。

在文章的最后，为了帮你把数据中台诞生的大事件串联起来，我做了一张时间图，在这个时间线里，你可以很清晰地看到数据中台诞生的前期、中期，和后期的大事件，这样可以帮你更清晰的掌握数据中台背景。

思考时间

在这节课快要结束时，我给你留一个发散性的思考题：如果说数据中台是大数据的下一站，那数据中台的下一站是什么？这个话题很有趣，欢迎你大开“脑洞”，在留言区与我分享。

最后，感谢你的阅读，如果这篇文章让你有所收获，也欢迎你将它分享给更多的朋友。

论文链接：

- 《The Google File System》
- 《MapReduce：Simplified Data Processing on Large Clusters》
- 《Bigtable：A Distributed Storage System for Structed Data》

© 版权归极客邦科技所有，未经许可不得传播售卖。 页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇　开篇词 | 数据中台，是陷阱？还是金钥匙？

下一篇　02 | 关键抉择： 到底什么样的企业应该建数据中台？

精选留言 (16)

写留言

西蒙 置顶

2020-03-31

数据中台的下一站是AI中台。

展开

作者回复: 机器学习是数据中台之上一个重要的应用领域。不是有句话是这么说的，数据和特征决定了机器学习的上限，算法和模型只是在无限的逼近这个上限。

我来谈谈数据中台的下一站到底是什么吧

目前数据中台的主要应用领域还是数据智能领域，所以我们就先不延申到机器学习，深度学习，安全、推荐等领域。

1. 实时数据中台，实现批流一体。
2. 云上数据中台，全面拥抱K8S，实现在线、离线混合部署，进一步提高资源利用率。

3. 智能元数据管理+增强分析，降低数据分析的门槛，进一步释放数据智能
4. 自动化代码构建，通过拖拉拽，自动化生成ETL代码的构建，进一步释放数据研发的效能，甚至让我们的非技术人员都可以完成简单的数据加工。
5. 数据产品的时代，面向各种行业的数据产品全面涌现，并且和中台系统联动，比如基于指标的可分析维度，自动进行指标的业务诊断等等。

我会在我们专栏结束的最后一篇，为大家在详细的展开这些，这里先抛砖引玉，提供给大家思考。

2

李跃爱学习
2020-03-30

老师讲得非常透彻，值回票价了

1. 传统数据仓库，第一次明确了数据分析的应用场景应该用单独的解决方案去实现，不再依赖于业务的数据库。在模型设计上，提出了数据仓库模型设计的方法论，为后来数据分析的大规模应用奠定了基础。...
展开

作者回复: 谢谢，前三篇主要还是从理论上来帮大家对齐概念，精华在实现篇中，有走过的坑，和避坑秘诀，继续看呦~ 欢迎多讨论。

6

Kǎfkā²⁰²⁰
2020-03-30

数据中台试图解决数据共享的问题，特别企业内部。再下一步，脑洞一下，我觉得数据要成为能在全社会共享的资源，像水和电那样，无处不在，按需索取。成为一种推动社会发展的信息能源。不妨叫数据能源?

作者回复: 想法太超前，数据是企业的核心资产，微信可能共享数据给其他企业嘛?

1

5

leslie
2020-03-31

还是追溯本源吧：OLAP其实是在OLTP之后出现的，Big Data的概念又是再次之后，中台

体现分析时还有共享，hadoop的资源损耗国内DB界一直饱有争议。举个例子：不同类型数据系统之间其实都是明显的有相互学习和继承的影子。
脑洞不用打开：中台体现的数据的集中、大数据体现的分散且快；那么下一步将是再次的分散。这就如同人的行走：不可能全是之路，曾经看到过一本资料中提出岛与海的概念-...
展开

作者回复: “为了中台而中台就失去了其原本的意义”

这句话说的好，数据中台要从解决问题的角度入手，不能解决问题，其实建中台意义不大。

2

Stefan
2020-03-31

老师，金博尔的建模设计方法的例子里面没有提到商品交易的信息，只有用户余额和库存事实表，这是为什么呢?
展开

作者回复: 在kimball建模中，只有事实和维度， 交易过程中涉及账户余额和库存，这都属于事实，而维度就是商品。因为从分析的角度，我们只关心事实在不同维度下的结果。而在商品交易中，我们需要分析的交易金额和账户余额，这都呈现在用户账户余额事实表中，只要在这个表中再关联商品的维度，就可以按照商品来分析，一个商品的交易金额。

比如，我们的余额事实表中，再增加一列代表商品的ID，每一个交易，都占用一行，代表交易完成后，用户余额和交易金额的数据。这样你就能实现从商品维度来分析交易金额的需求了。

1

邢爱明
2020-03-30

请教老师一个问题，传统企业是数据容量和类型是比较固定的，以前也有了数据仓库的应用，在什么场景下需要演建设数据中台?
如果要建，是否必须基于hadoop类的大数据平台?

作者回复: 数据中台一定是构建在数据湖之上的，这是与传统数据仓库，基于Oracle去构建，有本质的区别的。

原先数据仓库只能支持一些简单的报表，对数据的加工和处理能力都非常的有限，根本就没有办法支撑大规模的数据应用场景。如果你想实现数字化转型，真正让数据能够走入业务，让我们的业务人员每天看数据工作，那你就要构建数据中台了。

比如我们有一个零售的客户，他们构建了数据中台，然后在上面做了门店管家，数据应用，现在他们全国2W多家门店，每天都有大量的店员在看，有哪些商品卖的好，哪些卖的不好，哪些商品库存比较大，客户比较偏好买哪些商品，然后就会有针对性的推荐商品，或者调整商品的摆位。大幅度提高了门店的营业收入，这就是一个典型的数据中台的应用。

1

炒鸡辣鸡
2020-03-30

数据中台的下一站应该是对数据中台的可视化操作技术。产品经理直接拖拉数据中台中的组件即可以构造出用户满意的产品。这里面包涵的技术还是比较难的。

展开

作者回复: 虽然你的描述有点超前，但是方向是对的，我称之为自动化代码构建的能力，就是对于一些分析师、运营，我们也可以具备一些数据加工的能力。

1

幸运草
2020-03-30

我认为数据中台的下一站是加入深度学习的AI智能推荐和预测。

作者回复: 机器学习、深度学习，可以作为数据中台的一个应用领域，一般也是先有数据中台，才有机器学习，你这个回答，我认为也没毛病。

但是就数据这个领域而言，我给出的答案已经在置顶贴中，欢迎讨论~

1 1

希望
2020-03-31

数据中台解决的数据共享和数据服务的问题，下一站是解决更高层面和更细层面的问题，最终体现为智能化

陈
2020-03-31

数据中台的后面应该是基于区块链治理，把数据资产化，保护隐私的可信数据交换流通平台

陈
2020-03-31

数据中台概念有点类似数据虚拟化概念

展开

飞鱼
2020-03-31

老师讲的通俗易懂。

展开

作者回复: 感谢阅读~

xy
2020-03-31

郭老师还是想跟您探讨一下数据中台和数据仓库的区别。中台如何能够赋能应用？如何实现数据共享？

作者回复: 你好， 数据中台相比传统数据仓库，首先它构建于数据湖之上，其次它强调数据的复用性，强调数据只加工一次，避免烟囱式的开发模式，最后，它强调数据通过服务化的方式，实现数据应用和数据中台数据的打通。

数据中台与数据仓库的联系，可以认为数据中台继承了数据仓库中维度建模的设计方法。

中台如何赋能数据应用？
从效率、质量和成本三个维度。首先来说效率，数据中台在数据研发效率、数据发现效率、取数效率和数据分析效率都得到全面的提升，原先一个数据应用需求，最少需要一周研发时间，而现在可以缩短到2-3天，这样就大幅提高了数据应用的迭代速度，使得数据应用的迭代速度不再受制于数据开发的速度。当然还有质量，成本，你继续沿着我们的课程往下看，我会一一结合例子详细的介绍。

如何实现数据的共享？

我觉得这个问题，分为两个，一个是数据的共享，一个是接口的共享。原先我们存在很多分散的

小数仓，同一个小数仓内，也是按照应用，呈现烟囱式的数据加工。但是现在我们有了一个统一的数据中台，构建的是一个统一的数据中台，所以我们要实现模型的复用，强调相同粒度的事实只加工一次。并且在模型设计的度量中，我们也引入了模型引用系数，作为衡量模型复用的度量，以此推动大家尽量复用数据。另外数据复用的前提，还依赖高效的数据发现，得快速找到数据，理解数据，这与元数据的管理息息相关。在实现篇中，我都会一一讲解。接口层的复用，是通过数据服务实现的，我在数据服务章节也会详细介绍。

别急，慢慢看哈。

宋胖胖
2020-03-31

来学习，听过数据湖一直也不知道是什么，现在知道 了。数据湖（Data Lake）是一个以原始格式存储数据的存储库或系统。

展开

作者回复: 对的，你可以把Hadoop 理解为一个数据湖。

Max
2020-03-30

下一站是数据服务自动化，智能化。
不如我们起名叫“数据智能”。

作者回复: 真会蹭热点，哈哈~

数据智能，我认为代表着我们数据产品，数据应用全面爆发，深入业务，这才是真正的DT时代。

狼迹
2020-03-30

十年前学数据仓库，十年后折腾数据中台

展开

作者回复: 是不是觉得物是人非，又有种似曾相识?